

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re U.S. Patent Application of)
WATANABE et al.)
Application Number: To be assigned)
Filed: Concurrently Herewith)
For: DATA DISTRIBUTION METHOD, DATA SEARCH)
METHOD, AND DATA SEARCH SYSTEM)
Attorney Docket No. HIRA.0131)

Honorable Assistant Commissioner
for Patents
Washington, D.C. 20231

**REQUEST FOR PRIORITY
UNDER 35 U.S.C. § 119
AND THE INTERNATIONAL CONVENTION**

Sir:

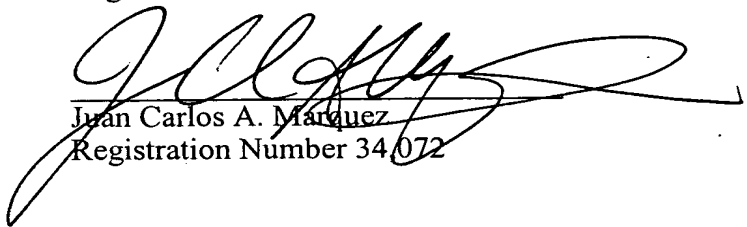
In the matter of the above-captioned application for a United States patent, notice is hereby given that the Applicant claims the priority date of November 27, 2002, the filing date of the corresponding Japanese patent application 2002-344452.

The certified copy of corresponding Japanese patent application 2002-344452 is being submitted herewith. Acknowledgment of receipt of the certified copy is respectfully requested in due course.

Respectfully submitted,

Stanley P. Fisher
Registration Number 24,344

REED SMITH LLP
3110 Fairview Park Drive
Suite 1400
Falls Church, Virginia 22042
(703) 641-4200



Juan Carlos A. Marquez
Registration Number 34,072

November 25, 2003

(Translation)

PATENT OFFICE
JAPANESE GOVERNMENT

This is to certify that the annexed is a true copy of
the following application as filed with this Office.

Date of Application: November 27, 2002

Application Number: Japanese Patent Application
No. 2002-344452

Applicant(s): Hitachi Software Engineering Co., Ltd.

October 8, 2003

Commissioner,
Patent Office

Yasuo IMAI (seal)

Certificate No. 2003-3083102

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 2002年11月27日
Date of Application:

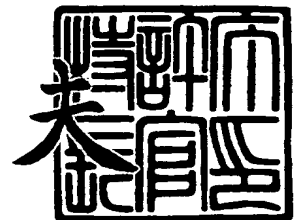
出願番号 特願2002-344452
Application Number:
[ST. 10/C]: [JP 2002-344452]

出願人 日立ソフトウェアエンジニアリング株式会社
Applicant(s):

2003年10月 8日

特許庁長官
Commissioner,
Japan Patent Office

今井 康



【書類名】 特許願

【整理番号】 14B009

【提出日】 平成14年11月27日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 17/30

【発明の名称】 データ配信方法、データ検索方法及びデータ検索システム

【請求項の数】 7

【発明者】

 【住所又は居所】 東京都品川区東品川4丁目12番7号 日立ソフトウェアエンジニアリング株式会社内

 【氏名】 渡辺 恒彦

【発明者】

 【住所又は居所】 東京都品川区東品川4丁目12番7号 日立ソフトウェアエンジニアリング株式会社内

 【氏名】 吉井 淳治

【発明者】

 【住所又は居所】 東京都品川区東品川4丁目12番7号 日立ソフトウェアエンジニアリング株式会社内

 【氏名】 水沼 貞

【発明者】

 【住所又は居所】 東京都品川区東品川4丁目12番7号 日立ソフトウェアエンジニアリング株式会社内

 【氏名】 峰崎 雄一

【発明者】

 【住所又は居所】 東京都品川区東品川4丁目12番7号 日立ソフトウェアエンジニアリング株式会社内

 【氏名】 小倉 文寿

【発明者】

【住所又は居所】 東京都品川区東品川 4 丁目 1 2 番 7 号 日立ソフトウェアエンジニアリング株式会社内

【氏名】 山本 圭介

【発明者】

【住所又は居所】 東京都品川区東品川 4 丁目 1 2 番 7 号 日立ソフトウェアエンジニアリング株式会社内

【氏名】 永井 健夫

【特許出願人】

【識別番号】 000233055

【氏名又は名称】 日立ソフトウェアエンジニアリング株式会社

【代理人】

【識別番号】 100091096

【弁理士】

【氏名又は名称】 平木 祐輔

【選任した代理人】

【識別番号】 100102576

【弁理士】

【氏名又は名称】 渡辺 敏章

【選任した代理人】

【識別番号】 100105463

【弁理士】

【氏名又は名称】 関谷 三男

【手数料の表示】

【予納台帳番号】 015244

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9722155

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 データ配信方法、データ検索方法及びデータ検索システム

【特許請求の範囲】

【請求項 1】 生体物質に関する情報を格納している複数のデータベースからデータをダウンロードするステップと、

前記ダウンロードしたデータから、インデックスとして、2つのデータベースのデータ間のリンクを表す情報、各データの詳細説明、及びホモロジー検索用の配列データを抽出するステップと、

抽出したインデックスを配信するステップとを含むことを特徴とするデータ配信方法。

【請求項 2】 生体物質に関する情報を格納している複数のデータベースからデータをダウンロードするステップと、

前記ダウンロードしたデータから、2つのデータベースのデータ間のリンクを表す情報を抽出するステップと、

検索キーとして、開始データベース名、ターゲットデータベース名、及び前記開始データベース中のデータ I Dを受け付けるステップと、

予め設定された複数のデータベース間におけるリンクの順序を表す情報を参照し、前記受け付けた開始データベース中のデータ I Dを起点として、前記抽出したデータ間のリンクのうち前記予め設定されたデータベース間におけるリンクの順序に適合するリンクをたどって前記ターゲットデータベースのデータ I Dを取得するステップと、

取得したターゲットデータベースのデータ I Dを表示するステップとを含むことを特徴とするデータ検索方法。

【請求項 3】 生体物質に関する情報を格納している複数のデータベースからデータをダウンロードするステップと、

前記ダウンロードしたデータから、2つのデータベースのデータ間のリンクを表す情報及びホモロジー検索用の配列データを抽出するステップと、

検索キーとして、開始データベース名、ターゲットデータベース名、及び入力配列データを受け付けるステップと、

前記入力配列データで前記開始データベースのホモロジー検索用配列データをホモロジー検索するステップと、

予め設定されたデータベース間におけるリンクの順序を表す情報を参照し、前記ホモロジー検索で求められた前記開始データベース中のデータ I D を起点として、前記抽出したデータ間のリンクのうち前記予め設定されたデータベース間におけるリンクの順序に適合するリンクをたどって前記ターゲットデータベースの対応するデータ I D を取得するステップと、

取得したターゲットデータベースのデータ I D を表示するステップとを含むことを特徴とするデータ検索方法。

【請求項 4】 生体物質に関する情報を格納している複数のデータベースから 2 つのデータベースのデータ間のリンクを表す情報を収集したインデックスデータを用意するステップと、

前記複数のデータベース間におけるリンクの順序を定めたテーブルを用意するステップと、

検索キーとして、開始データベース名、ターゲットデータベース名、及び前記開始データベース中のデータ I D を受け付けるステップと、

前記受け付けた開始データベース中のデータ I D を起点として、前記データ間のリンクのうち前記データベース間におけるリンクの順序に適合するリンクをたどって前記ターゲットデータベースの対応するデータ I D を取得するステップと、

取得したターゲットデータベースのデータ I D を表示するステップとを含むことを特徴とするデータ検索方法。

【請求項 5】 生体物質に関する情報を格納している複数のデータベースから 2 つのデータベースのデータ間のリンクを表す情報及びホモロジー検索用の配列データを収集したインデックスデータを用意するステップと、

前記複数のデータベース間におけるリンクの順序を定めたテーブルを用意するステップと、

検索キーとして、開始データベース名、ターゲットデータベース名、及び入力配列データを受け付けるステップと、

前記入力配列データで前記開始データベースのホモロジー検索用配列データをホモロジー検索するステップと、

前記ホモロジー検索で求められた前記開始データベース中のデータ I D を起点として、前記データ間のリンクのうち前記複数のデータベース間におけるリンクの順序に適合するリンクをたどって前記ターゲットデータベースの対応するデータ I D を取得するステップと、

取得したターゲットデータベースのデータ I D を表示するステップとを含むことを特徴とするデータ検索方法。

【請求項 6】 生体物質に関する情報を格納している複数のデータベースから 2 つのデータベースのデータ間のリンクを表す情報を収集したインデックスデータと、

前記複数のデータベース間におけるリンクの順序を定めたテーブルと、

検索キーとして、開始データベース名、ターゲットデータベース名、及び前記開始データベース中のデータ I D を受け付ける入力部と、

前記受け付けた開始データベース中のデータ I D を起点として、前記データ間のリンクのうち前記データベース間におけるリンクの順序に適合するリンクをたどって前記ターゲットデータベースの対応するデータ I D を取得する検索部と、

取得したターゲットデータベースのデータ I D を表示する表示部とを含むことを特徴とするデータ検索システム。

【請求項 7】 生体物質に関する情報を格納している複数のデータベースから 2 つのデータベースのデータ間のリンクを表す情報及びホモロジー検索用の配列データを収集したインデックスデータと、

前記複数のデータベース間におけるリンクの順序を定めたテーブルと、

検索キーとして、開始データベース名、ターゲットデータベース名、及び入力配列データを受け付ける入力部と、

前記入力配列データで前記開始データベースのホモロジー検索用配列データをホモロジー検索する第 1 検索部と、

前記ホモロジー検索で求められた前記開始データベース中のデータ I D を起点として、前記データ間のリンクのうち前記複数のデータベース間におけるリンク

の順序に適合するリンクをたどって前記ターゲットデータベースの対応するデータ I D を取得する第 2 検索部と、

取得したターゲットデータベースのデータ I D を表示する表示部とを含むことを特徴とするデータ検索システム。

【発明の詳細な説明】

【 0 0 0 1 】

【発明の属する技術分野】

本発明は、塩基配列、タンパク質配列などの生体物質に関する情報を格納する複数のデータベースを関連付けることにより、データベース間のつながりをまとめ、これより情報を検索する方法に関する。

【 0 0 0 2 】

【従来の技術】

生体物質に関する情報を蓄積したデータベースは世界中に存在し、Web上で公開されている。生物の研究者はこれらのデータベースを活用し、自分の研究に役立てている（非特許文献 1 参照）。遺伝子情報、タンパク質情報に関する公開データベースは、データベース固有の登録番号（以下、I D という）を持ち、これをそのデータベースが格納している遺伝子、タンパク質に割り当てている場合が多い。これまで研究者が自分のデータを公開データベースに対し検索し、データベース中のデータを取得する場合、何らかの手段を用いて自分のデータとデータベース固有の I D との関連付けを行う必要があった。その最も一般的な方法は、研究者の持つ塩基配列、タンパク質配列とデータベースに格納してある塩基配列、タンパク質配列のホモロジー検索を行い、対応付けを行う方法である。

【 0 0 0 3 】

これには大きく分けて 2 つの方法がある。1 つは自分のデータを公開データベースに対して一つ一つ Web 検索する方法である。もう 1 つは、インターネットを介して検索することによる情報漏洩を防止するため、自分の施設内に Web 上のデータベースのデータを一つ一つダウンロードし、これに対して検索する方法である。図 2 1 は、Web 上のデータベースのデータをダウンロードして検索する従来のシステムの模式図である。ユーザ 2 1 8 は、公開データベース 2 1 1 よりイン

ターネット 2 1 2 を介してユーザの施設 2 1 7 にファイル 2 1 9 を一つ一つダウンロードする。その後、ユーザ 2 1 8 は、ダウンロードしたファイル 2 1 9 に対して検索を行う。

【非特許文献 1】

Baxebanis, A.D:Nucl.Acids Res.,28:1-10,2000, "Genetics Databases"
(Bishop M.J ed.), Academic Press, Cambridge,1999

【0 0 0 4】

【発明が解決しようとする課題】

従来は、研究者が一度に扱うデータ数は 1 個から 1 0 個程度であったので、一つ一つWeb検索して情報検索することも可能だったが、近年の技術革新により数百から数千のデータを扱うようになり、一つ一つ検索するのは非常に煩雑な作業になった。また、複数の公開データベースを検索すると、不要なデータが検索結果として得られ、これより自分の必要な情報を再度抽出しなければならなかった。さらに、世界中にはたくさんのデータベースがあり、そのデータベースが自分に必要なものかどうか検討する必要があった。データベースの中には複数の生物種（ヒト、マウス、イネ等）が含まれている場合があり、ある生物種に関するデータをさまざまなデータベースから網羅的に取得するようなシステムはなかった。同様に、データの種別（DNA, mRNA, EST）に取得するようなシステムはなかった。

【0 0 0 5】

また、ユーザの施設内に複数の公開データベースからデータを一つ一つダウンロードする場合、ダウンロードするデータ量が多ければ、長時間かかり、ダウンロードの最中に回線が中断されてしまうという問題があった。また、ダウンロードのために長時間回線が占有されてしまうという問題があった。これに加え、現在は生物関連情報が急激に増加しており、今後のダウンロード作業はさらに手間取るようになることが考えられる。また、公開データベースの情報はそれぞれのデータベース管理者により管理されているため、生物の研究者がそれぞれの公開データベースの更新時期、現在のデータ数等を常に把握するのは困難であった。

【0 0 0 6】

また、データベース間にはさまざまなリンクが存在する。従って、データを検索する場合、複数のリンクをたどって検索を行っていた。例えば図 22 に示す通り、データベース A のデータに対応するデータベース D のデータを取得する場合、データベース B を経由するルートとデータベース C を経由するルートがある。データベース A の遺伝子 A 1 に対応するデータベース B のデータは B 1, B 2 であり、これに対応するデータベース D のデータは D 1, D 2 である。また、遺伝子 A 1 に対応するデータベース C のデータは C 1、これに対応するデータベース D のデータは D 3 である。この例の場合には、データベース A の遺伝子 A 1 に対応するデータベース D のデータが D 1, D 2, D 3 と 3 つあることになり、ユーザはどれが正しいデータが再度検証する必要がある。

【0007】

本発明は、このような生体物質の情報に関するデータベース検索の現状に鑑み、ネットワーク上のデータベースのデータを容易に検索できる方法及びシステムを提供することを目的とする。

【0008】

【課題を解決するための手段】

本発明においては、複数のデータベースより必要な情報を抽出してインデックスを作成し、これを配信する方法をとる。これにより、ユーザは必要な情報のみを得られるようになる。複数のデータを 1 つのインデックスにまとめてあるため、データ量が軽減され、データセンタからユーザの施設内へのダウンロードもスムーズに行われ、ダウンロードのために長時間回線が占有されてしまうという問題もない。また、データセンタでデータベースの更新、フォーマットの変更などに関して一括して対応できるため、ユーザはそれらの煩雑な作業から解放される。なお、情報の漏洩等の心配がない場合には、インデックスをユーザの施設内にダウンロードせずに、データセンタにおかれたインデックスに直接アクセスして検索を行ってもよい。

【0009】

すなわち、本発明によるデータ配信方法は、生体物質に関する情報を格納している複数のデータベースからデータをダウンロードするステップと、ダウンロー

ドしたデータから、インデックスとして、2つのデータベースのデータ間のリンクを表す情報、各データの詳細説明、及びホモロジー検索用の配列データを抽出するステップと、抽出したインデックスを配信するステップとを含むことを特徴とする。

【0010】

本発明によるデータ検索方法は、生体物質に関する情報を格納している複数のデータベースからデータをダウンロードするステップと、ダウンロードしたデータから、2つのデータベースのデータ間のリンクを表す情報を抽出するステップと、検索キーとして、開始データベース名、ターゲットデータベース名、及び開始データベース中のデータIDを受け付けるステップと、予め設定された複数のデータベース間におけるリンクの順序を表す情報を参照し、受け付けた開始データベース中のデータIDを起点として、前記抽出したデータ間のリンクのうち予め設定されたデータベース間におけるリンクの順序に適合するリンクをたどってターゲットデータベースのデータIDを取得するステップと、取得したターゲットデータベースのデータIDを表示するステップとを含むことを特徴とする。

【0011】

本発明によるデータ検索方法は、また、生体物質に関する情報を格納している複数のデータベースからデータをダウンロードするステップと、ダウンロードしたデータから、2つのデータベースのデータ間のリンクを表す情報及びホモロジー検索用の配列データを抽出するステップと、検索キーとして、開始データベース名、ターゲットデータベース名、及び入力配列データを受け付けるステップと、入力配列データで開始データベースのホモロジー検索用配列データをホモロジー検索するステップと、予め設定されたデータベース間におけるリンクの順序を表す情報を参照し、ホモロジー検索で求められた開始データベース中のデータIDを起点として、前記抽出したデータ間のリンクのうち予め設定されたデータベース間におけるリンクの順序に適合するリンクをたどってターゲットデータベースの対応するデータIDを取得するステップと、取得したターゲットデータベースのデータIDを表示するステップとを含むことを特徴とする。

【0012】

本発明によるデータ検索方法は、また、生体物質に関する情報を格納している複数のデータベースから2つのデータベースのデータ間のリンクを表す情報を収集したインデックスデータを用意するステップと、前記複数のデータベース間におけるリンクの順序を定めたテーブルを用意するステップと、検索キーとして、開始データベース名、ターゲットデータベース名、及び開始データベース中のデータIDを受け付けるステップと、受け付けた開始データベース中のデータIDを起点として、前記データ間のリンクのうちデータベース間におけるリンクの順序に適合するリンクをたどってターゲットデータベースの対応するデータIDを取得するステップと、取得したターゲットデータベースのデータIDを表示するステップとを含むことを特徴とする。

【0013】

本発明によるデータ検索方法は、また、生体物質に関する情報を格納している複数のデータベースから2つのデータベースのデータ間のリンクを表す情報及びホモロジー検索用の配列データを収集したインデックスデータを用意するステップと、前記複数のデータベース間におけるリンクの順序を定めたテーブルを用意するステップと、検索キーとして、開始データベース名、ターゲットデータベース名、及び入力配列データを受け付けるステップと、入力配列データで開始データベースのホモロジー検索用配列データをホモロジー検索するステップと、ホモロジー検索で求められた開始データベース中のデータIDを起点として、前記データ間のリンクのうち複数のデータベース間におけるリンクの順序に適合するリンクをたどってターゲットデータベースの対応するデータIDを取得するステップと、取得したターゲットデータベースのデータIDを表示するステップとを含むことを特徴とする。

【0014】

本発明によるデータ検索システムは、生体物質に関する情報を格納している複数のデータベースから2つのデータベースのデータ間のリンクを表す情報を収集したインデックスデータと、前記複数のデータベース間におけるリンクの順序を定めたテーブルと、検索キーとして、開始データベース名、ターゲットデータベース名、及び開始データベース中のデータIDを受け付ける入力部と、受け付け

た開始データベース中のデータIDを起点として、前記データ間のリンクのうちデータベース間におけるリンクの順序に適合するリンクをたどってターゲットデータベースの対応するデータIDを取得する検索部と、取得したターゲットデータベースのデータIDを表示する表示部とを含むことを特徴とする。

【0015】

本発明によるデータ検索システムは、また、生体物質に関する情報を格納している複数のデータベースから2つのデータベースのデータ間のリンクを表す情報及びホモロジー検索用の配列データを収集したインデックスデータと、前記複数のデータベース間におけるリンクの順序を定めたテーブルと、検索キーとして、開始データベース名、ターゲットデータベース名、及び入力配列データを受け付ける入力部と、入力配列データで開始データベースのホモロジー検索用配列データをホモロジー検索する第1検索部と、ホモロジー検索で求められた開始データベース中のデータIDを起点として、前記データ間のリンクのうち複数のデータベース間におけるリンクの順序に適合するリンクをたどってターゲットデータベースの対応するデータIDを取得する第2検索部と、取得したターゲットデータベースのデータIDを表示する表示部とを含むことを特徴とする。

【0016】

本発明によると、インデックスに対して、数千のデータを一括検索できるようになる。また、ネットワークを構築するときに用いるデータベースを生物種別（ヒト、マウス、イネ等）、データの種別（DNA、mRNA、EST）等に分類整理しておくことにより、ユーザは目的に合わせたデータを取得できるようになる。また、複数のデータベース間におけるリンクの順序を定めたテーブル等を用意しておき、そこに定められたルートに従って複数のデータベースのリンクをたどることにより、ノイズの少ない検索結果を得ることができる。

【0017】

【発明の実施の形態】

以下、図面を参照して本発明の実施の形態について説明する。

図1は、本発明による生体物質情報検索システムの仕組みの一例を示す概念図である。公開データベースや商用データベース11のデータは、インターネット

1 2 を介してデータセンタ 1 3 にダウンロードされる。データセンタ 1 3 では、ダウンロードされたデータからインデックス 1 5 を作成する。作成されたインデックス 1 5 はユーザの施設 1 7 に対して配信され（インデックス 1 6）、ユーザ 1 8 は配信されたインデックス 1 6 に対して検索を行う。

【0 0 1 8】

インデックスには、異なるデータベースに含まれるデータ間の対応関係を表すリンク情報、各データの詳細説明、ホモロジー検索用データが含まれる。各データの詳細説明とは、データベースのエントリ一つ一つに格納されているエントリの詳細説明である。ホモロジー検索用データとは、データベースに含まれている塩基配列やタンパク質配列などの配列情報である。ユーザは、自分の有する塩基配列もしくはタンパク質配列と、目的となる公開データベースのデータの塩基配列もしくはタンパク質配列との間でホモロジー検索を行う。ホモロジー検索を行うソフトウェアには通常BLASTが用いられるため、ホモロジー検索用データはファスタ形式の配列データをBLAST用にフォーマットしたものを用いる。

【0 0 1 9】

なお、ネットワークを構築するときに用いるデータベースは生物種別（ヒト、マウス、イネ等）、データの種別（DNA, mRNA, EST）に分類整理しておくことにより、ユーザは目的に合わせたデータを取得できるようになる。

【0 0 2 0】

図 2 は、生体物質情報を格納する複数のデータベースから、情報を検索するためのインデックスを作成する手順を示すフローチャートである。

まずステップ 1 1 において、公的データベースや商用データベース等の公開されているデータベースからデータセンタにデータをダウンロードする。次に、ステップ 1 2 において、ダウンロードしたデータから、リンク情報、ホモロジー検索用データ及びそれぞれの ID の詳細説明を自動抽出する。このとき、ホモロジー検索用データは、インデックスに登録するデータベースのうちで、配列情報が存在するすべてのデータベースについて取得する。また、詳細情報は、インデックスに登録するすべてのデータベースについて取得する。最後に、ステップ 1 3 において、リンク情報、ホモロジー検索用データ、それぞれの ID の詳細説明を

まとめてユーザの施設に配信する。

【0021】

図3は、図2のステップ12におけるリンク情報作成の手順を説明する図である。図示した例では、データベースAはデータベースBとデータベースEに対応しており、データベースAのエントリであるA1に対して、データベースBのエントリB1とデータベースEのエントリE1が対応しており、これがデータベースAファイルに記述してある。従って、データベースAファイルよりそれぞれのIDを取り出し、データベースAのA1とデータベースBのB1をテーブル31に格納する。同様に、データベースAのエントリA1とデータベースEのエントリE1の対応が記述されており、これらを取り出しテーブル32に格納する。データベースBファイルにはデータベースBのエントリB1とデータベースCのエントリC1の対応が記述されており、これらを取り出しテーブル33に格納する。データベースCファイルにはデータベースCのエントリC1とデータベースDのエントリD1の対応が記述されており、これらを取り出しテーブル34に格納する。これらのテーブル31、33、34をつなぎ合わせることでよりテーブル35を作成できる。テーブル32とテーブル35を模式図で表すとリンク図36のようになる。

【0022】

図4は、リンク情報から得られるルートの他の例を示す図である。リンク情報としてデータベースに格納されているテーブルは、図3のテーブル31～34等示すように、2つのデータベースのIDが対応したものになっている。これより、図4に示す表41もしくは表42を作成する。これらの表をデータベース間の関係を示す模式図で表すとリンク図43のようになる。このリンク図43上で対応するデータをたどっていくことにより、例えばデータベースAのデータA1に対応するデータベースDのデータD1を検索することができる。

【0023】

ここで、データベースには他の種々のデータベースとのリンク情報が記載されており、リンクが錯綜することにより図22によって説明したような問題が生じることがある。そこで、本発明においては、それぞれのデータベース同士は決め

られたルール（順序）に従ってリンクするように、データベース間のリンクを制限する。データベース間のリンクの制限について以下に説明する。

【0024】

図5は、許容されるデータベース間のリンクのルート（順序）に関する情報を格納したルートテーブルの例を示す図である。「KeyDB」は検索の起点となるデータベース、「TargetDB」は「KeyDB」中のデータに対応するデータを求めたいデータベースである。公開データベース、商用データベース、個人のオリジナルデータのデータベース等からなるデータベースA、B、C、…には、そのデータベースのあるデータが他のデータベースのどのデータに対応するかを示すデータ間のリンクの情報が複数記述されている場合があり、種々のルートをたどってKeyDB中の指定データに対応するTargetDB中のデータを検索することが可能であるが、全てのリンク情報を利用すると、前述したようにノイズ情報を拾う可能性がある。そこで、KeyDBとTargetDBを指定すると、KeyDBからTargetDBに至るリンクのルート（順序）をルートテーブルによって一意に指定する。図示の例では、KeyDBがAでTargetDBがCの場合には、図5のルートテーブルを参照して、データベースA、データベースB、データベースCの順にリンクをたどってデータベースA中のデータに対応するデータベースC中のデータを検索する。同様に、KeyDBがBでTargetDBがDの場合には、図5のルートテーブルを参照して、データベースA、データベースB、データベースC、データベースDの順にリンクをたどってデータベースA中のデータに対応するデータベースD中のデータを検索する。

【0025】

図6は、ルートテーブルの内容をネットワーク表示した例を示す図である。データベース61と63とが対応していることを2つのデータベースを結ぶ線62が表している。いま、データベース61はデータベース63に格納されているデータを元に新たに作成されており、例えばデータベース61に格納されているデータAがデータベース63に格納されているデータBに対応しているとする。本発明では、このようなデータの起源に従ったリンク情報のみを利用し、例えばデータベース61に他のデータベース64とのリンク情報が格納されていても、そ

れは検索のためのリンク情報としては利用しない。このようにデータベース間のリンクを制限することにより、不要なデータの取得を制限することができる。

【0026】

図7は、データベース間のリンクを制限したことによる効果を説明する図であり、図22に対応する図である。

データベースAにデータベースBへのリンク情報とデータベースCへのリンク情報が記述されている場合、本発明では、より信頼性の高いデータベースAとデータベースCの間のリンク情報のみ利用し、データベースAとデータベースBの間のリンク情報は利用しない。その結果、データベースA中の遺伝子データA1に対応するデータベースDの遺伝子データD3を取得することができる。このようにデータベース間のリンクを制限することにより、図22に示したようなノイズとなる余分なデータの取得を制限し、適切なデータのみを取得することができるようになる。

【0027】

図8は、図2のステップ12におけるホモロジー検索用データの作成手順を示す図である。ここには、公開データベースからダウンロードしたファイル81から各エントリーのID83と配列データ84を抽出し、FASTA形式の配列データ85を格納したファイル82を作成する例を示している。

【0028】

図9は、詳細説明ファイルの作成手順を示す図である。ここには、公開データベースからダウンロードしたファイル91から各エントリーのID93とそのエントリに関する詳細説明94を抽出し、詳細説明ファイル92にIDと詳細説明の組95として格納する例を示している。

【0029】

図10は、インデックス情報の詳細について示す図である。データセンタ13において、インデックス情報（リンク情報101、詳細説明103、ホモロジー検索用データ106）を作成する。リンク情報101はネットワークに登録したデータベースより取得したリンク用テーブル102の形で保持する。詳細説明103は、ネットワークに登録したデータベースより取得した詳細説明用テーブル

104として保持する。リンク情報と詳細説明のそれぞれのテーブルをデータベース107に格納する。また、ファスタ形式のファイル105に対しBLASTで使用するようフォーマットし、ホモロジー検索用データ106を作成する。データセンタ13で作成したこれらのインデックス情報をユーザの施設17に作成する。この場合、データベース107の複製をレプリケーション処理により、ユーザの施設17のデータベース108に作成する。また、ホモロジー検索用データ106のコピー109を、ユーザの施設17に転送する。また、データベース間のリンクのルート（順序）に関する情報を格納したルートテーブル110のコピー111もユーザの施設17に転送される。

【0030】

図11は、本発明による生体物質情報検索の手順を示すフローチャートである。また、図12は、この検索方法を実現するための検索システムの概略構成図である。

【0031】

本発明による検索システムは、図10にて説明したリンク情報及び詳細説明を格納したデータベース124、ホモロジー検索用データ125、データベース間のリンク順序を記載したルートテーブル126、入力操作部127、検索結果を表示する表示部128、及び検索処理部121を備える。検索処理部121は、リンクをたどってID検索を行うID検索部122と、入力操作部から入力された配列データとホモロジー検索用データの間でホモロジー検索を行うホモロジー検索部123を有する。図13、図14はデータ検索時の入力インタフェースの例を示す説明図である。図13はデータベースのIDを検索する場合に用いる入力インタフェース、図14は塩基配列、タンパク質配列を検索する場合に用いる入力インタフェースである。

【0032】

最初に、ユーザデータのIDをネットワーク上のデータベースのIDに変換する検索方法及び検索システムについて説明する。

まず、図11のステップ21において、入力操作部127を操作してデータの入力を行う。例えば、図15の例に示すような入力データのファイルを図13に

示す画面の「File Upload」ボタン 132 で選択すると、図 13 のデータ入力フィールド 131 にデータがカンマ区切りで表示される。「Clear」ボタン 133 を押すと入力データがクリアされる。図 15 に示した入力データ例は、NCBI で公開している UniGene のデータを示したものである。

【0033】

図 11 のステップ 22 では、KeyDB、TergetDB の設定を行う。入力データと同じ ID をもつデータベースを図 13 の KeyDB リスト 134 から選択し、変換対象となるデータベースを図 13 の TergetDB リスト 135 で選択する。すると、ルートテーブル 126 を参照して、フィールド 136 に検索ルートが表示される。また、ボタン 137 を選択すると、ID ネットワークの全体図の図 6 が表示され、KeyDB と TergetDB を確認することができる。

【0034】

次に、ステップ 23 において検索開始ボタン 138 を押し、検索を開始する。ID 検索部 122 の検索プログラムは、指定された検索ルートをたどって入力された KeyDB のデータ ID に対応する TergetDB のデータ ID を検索する。

【0035】

次に、ステップ 24 に進み、検索結果の表示を行う。図 16 は、検索結果を表示する表示部 128 の表示画面例を示す図である。この図の例では、フィールド 161 に KeyDB である UniGene のエントリ 162 に対応する TergetDB の SWISS-PROT のエントリ 163 を示している。「Hit Count」166 には KeyDB のエントリ 162 に対応する TergetDB のエントリ 163 の数を表示している。KeyDB ボタンもしくは Terget DB ボタン 164 をクリックすることにより、図 17 に示すような詳細説明が表示される。また、「View Route」ボタン 165 をクリックすると図 6 に示すようなデータベース間の検索ルートを示した図が表示される。

【0036】

次に、ユーザの検索したい塩基配列もしくはタンパク質配列を ID ネットワーク上のデータベースの ID に変換する場合の例について説明する。

図 11 のステップ 21 において、入力操作部 127 から検索したい配列データの入力を行う。例えば、図 14 に示す入力画面の「File Upload」ボタン 146

をクリックし、図 1 8 に例示するような入力データのファイルを選択すると、入力画面のデータ入力フィールド 1 4 1 に、入力した配列データが表示される。「Clear」ボタンをクリックするとデータ入力フィールド 1 4 1 は空になる。

【0 0 3 7】

次に、ステップ 2 2 に進み、KeyDB、TargetDBの設定を行う。検索データに対して対応させたいデータベース（KeyDB）を図 1 4 に示す入力画面のDBリスト 1 4 9 で選択し、変換対象となるデータベース（TargetDB）を図 1 4 のTargetDBリスト 1 4 3 で選択する。KeyDBの設定の後、検索したい配列データとKeyDBとなるデータベースに格納されているデータが核酸配列かタンパク質配列かにより、プログラムリスト 1 4 2 から適当なBLAST手法を選択する。例えば、「blastn (DNA Query vs. DNA DB)」は核酸配列の検索データで核酸配列データベースをサーチする。「blastp (Protein Query vs. Protein DB)」はタンパク質配列のクエリーでタンパク質配列データベースをサーチする。「blastx (DNA Query vs. Protein DB)」は核酸配列のクエリーを 6 フレーム翻訳してタンパク質配列データベースをサーチする。「tblastn (Protein Query vs. DNA DB)」はタンパク質配列のクエリーで核酸配列データベースを動的に6フレームに翻訳しながらサーチする。また、BLAST検索の詳細なパラメータの設定を詳細オプション設定部 1 4 7 において行う。

【0 0 3 8】

「View Route」ボタン 1 4 4 を押すと、データベースネットワークの全体図である図 6 を表示し、KeyDBとTargetDBの位置を確認することができる。また、フィールド 1 4 8 にはルートテーブルに設定されている検索ルートが表示される。

【0 0 3 9】

次にステップ 2 3 に進み、検索開始ボタン 1 4 5 を押すと、検索を開始する。検索に当たっては、最初にホモロジー検索部 1 2 3 の検索プログラム（BLAST）が起動し、入力した配列データとKeyDBとして指定されたデータベースのホモロジー検索用データとの間でホモロジー検索を行い、候補データの I D を取得する。次に I D 検索部 1 2 2 の検索プログラムが起動し、ホモロジー検索によって取得したKeyDBの I D を起点として、ルートテーブルによって設定されたリンクの

ルートをたどってTargetDBの対応する I D 検索が行われる。

【 0 0 4 0 】

ステップ 2 4 では、検索結果を表示する。図 1 9 は、検索結果を表示する表示部の画面例を示す図である。図示した例では、フィールド 1 9 1 に、KeyDB (Nucleotide(EST)) の I D 1 9 1 に対応するTargetDB (SWISS-PROT) の I D 1 9 3 を示している。また「Hit Count」 1 9 7 に、KeyDBのNucleotide(EST)の I D に対応するTargetDBのSWISS-PROT の I D の数を示している。「KeyDB」ボタンもしくは「Target DB」ボタン 1 9 4 をクリックすることにより、図 1 7 に示したような詳細説明を表示させることができる。また「ViewAlignment」ボタン 1 9 5 をクリックすることにより、図 2 0 に示すようなホモロジー検索結果が表示される。図 2 0 の「E-value」とは期待値、「Score」とは相同性の値のことである (Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.)。最もScoreの高いデータの I D を検索キーとし、I D 検索を行う。

【 0 0 4 1 】

【発明の効果】

本発明によると、ネットワークのリンクをたどることにより、ネットワーク上のすべてのデータベースのデータを容易に検索できるようになる。

【図面の簡単な説明】

【図 1】

本発明による生体物質情報検索システムの仕組みの一例を示す概念図。

【図 2】

生体物質情報を格納する複数のデータベースから、情報を検索するためのインデックスを作成する手順を示すフローチャート。

【図 3】

リンク情報作成の手順を説明する図。

【図 4】

リンク情報から得られるルートの他の例を示す図。

【図 5】

データベース間のリンクのルート（順序）に関する情報を格納したルートテーブルの例を示す図。

【図 6】

ルートテーブルの内容をネットワーク表示した例を示す図。

【図 7】

データベース間のリンクを制限したことによる効果を説明する図。

【図 8】

ホモロジー検索用データの作成手順を示す図。

【図 9】

詳細説明ファイルの作成手順を示す図。

【図 1 0】

インデックス情報の詳細について示す図。

【図 1 1】

本発明による生体物質情報検索の手順を示すフローチャート。

【図 1 2】

本発明による検索システムの概略構成図。

【図 1 3】

データベースの I D を検索する場合に用いるインタフェースの例を示す図。

【図 1 4】

配列を検索する場合に用いるインタフェースの例を示す図。

【図 1 5】

入力データ例を示す図。

【図 1 6】

検索結果を表示する表示部の画面例を示す図。

【図 1 7】

詳細説明の表示例を示す図。

【図 1 8】

入力データのファイル例を示す図。

【図 1 9】

検索結果を表示する表示部の画面例を示す図。

【図 2 0】

ホモロジー検索結果の表示例を示す図。

【図 2 1】

Web上のデータベースのデータをダウンロードして検索する従来のシステムの模式図。

【図 2 2】

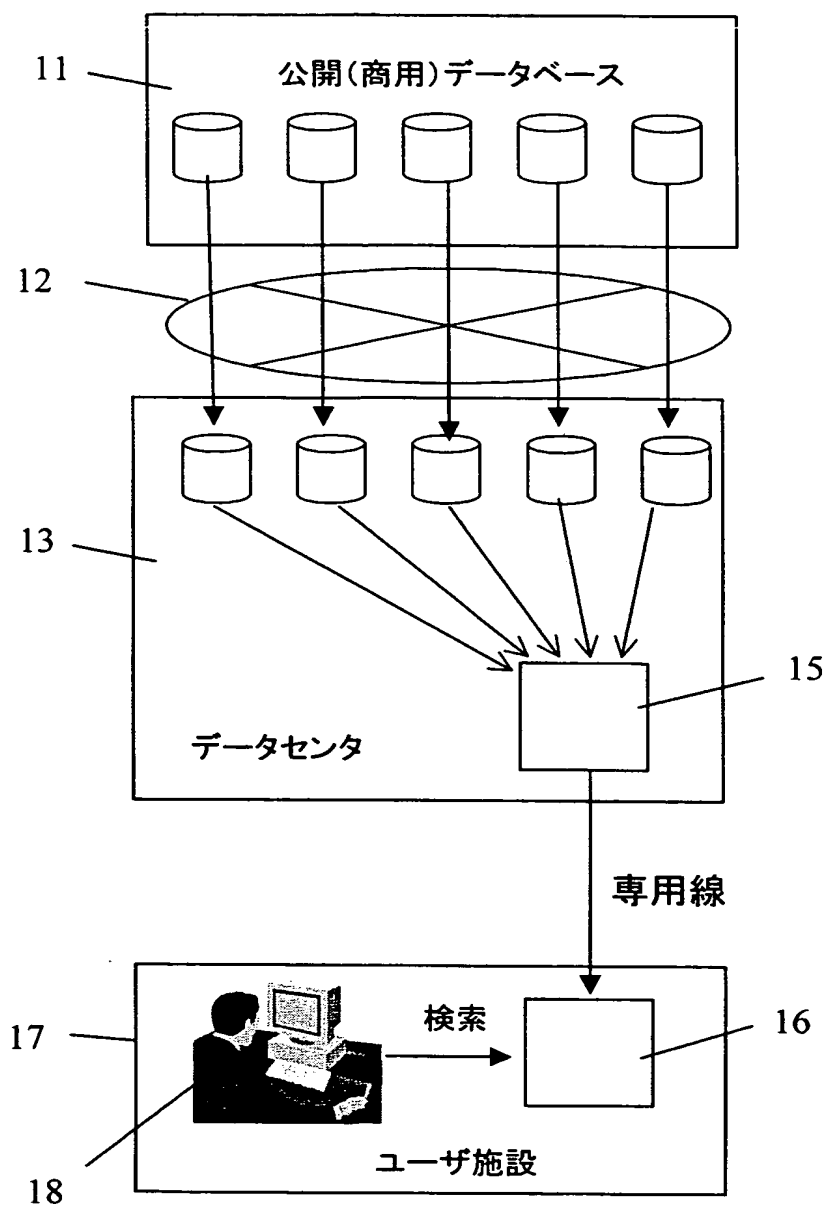
複数のリンクをたどって検索を行う場合の説明図。

【符号の説明】

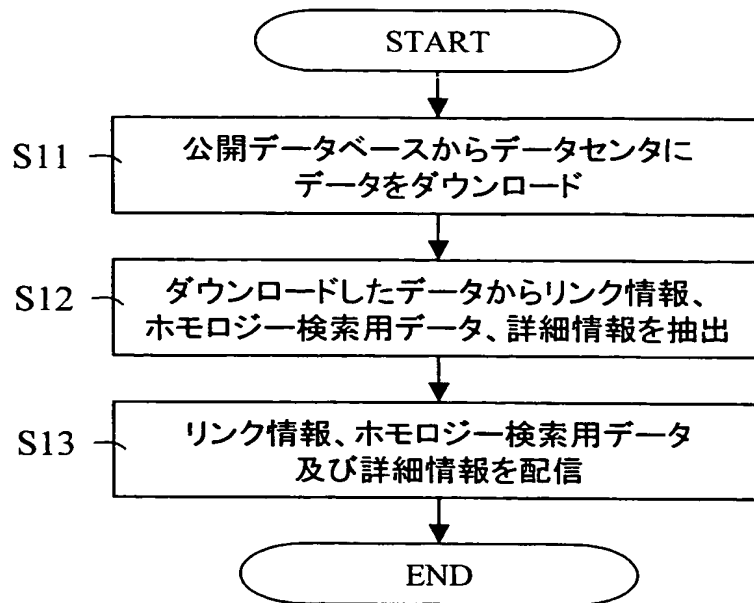
1 1…公開データベースあるいは商用データベース、1 2…インターネット、1 3…データセンタ、1 5…インデックス、1 6…配信されたインデックス、1 7…ユーザの施設、1 8…ユーザ、8 1…公開データベースからダウンロードしたファイル、8 5…FASTA形式の配列データ、9 2…詳細説明ファイル、1 0 1…リンク情報、1 0 3…詳細説明、1 0 6…ホモロジー検索用データ、1 0 7…データベース、1 2 1…検索処理部、1 2 2…ID検索部、1 2 3…ホモロジー検索部、1 2 4…リンク情報及び詳細説明を格納したデータベース、1 2 5…ホモロジー検索用データ、1 2 6…データベース間のリンク順序を記載したルートテーブル

【書類名】 図面

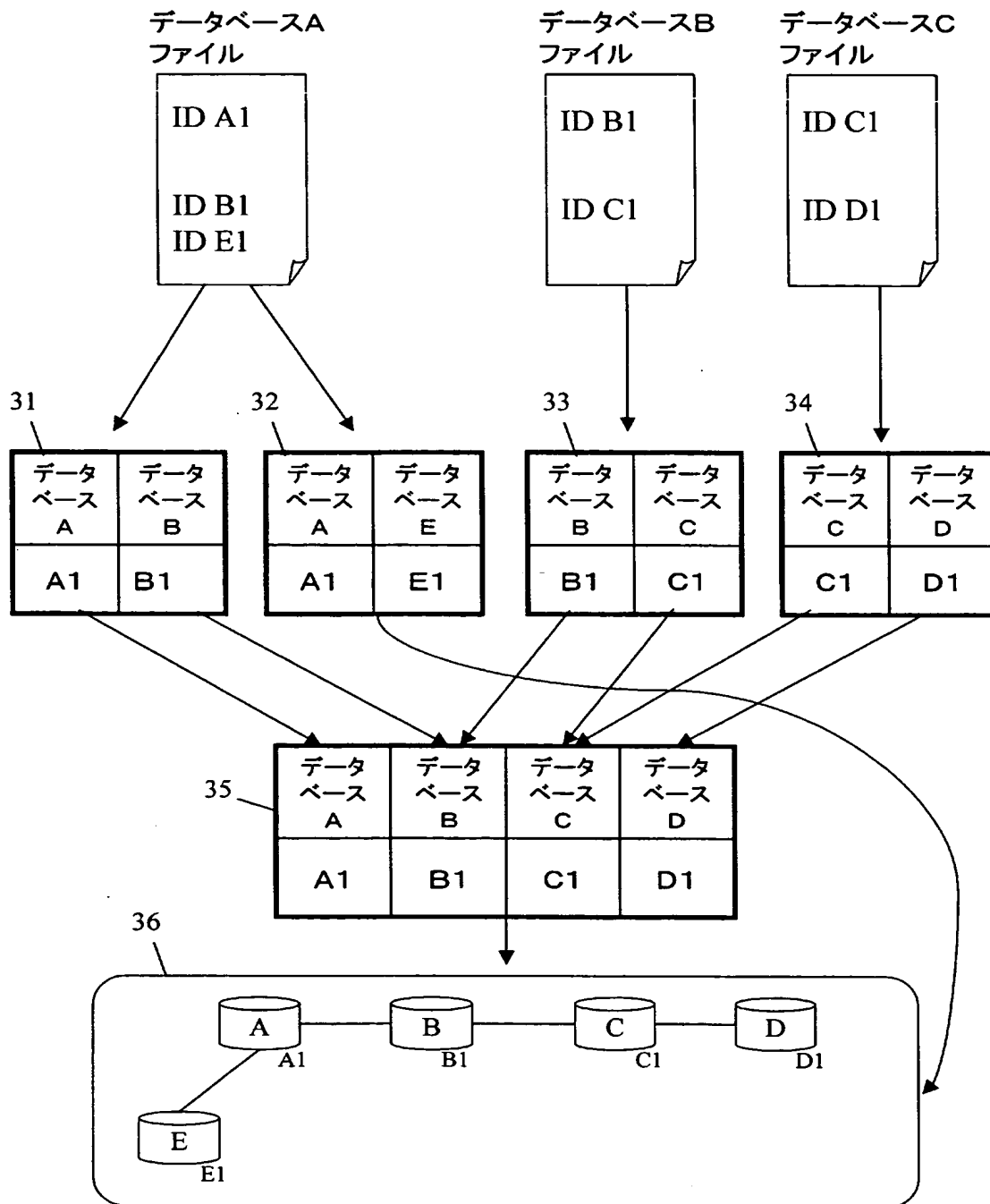
【図 1】



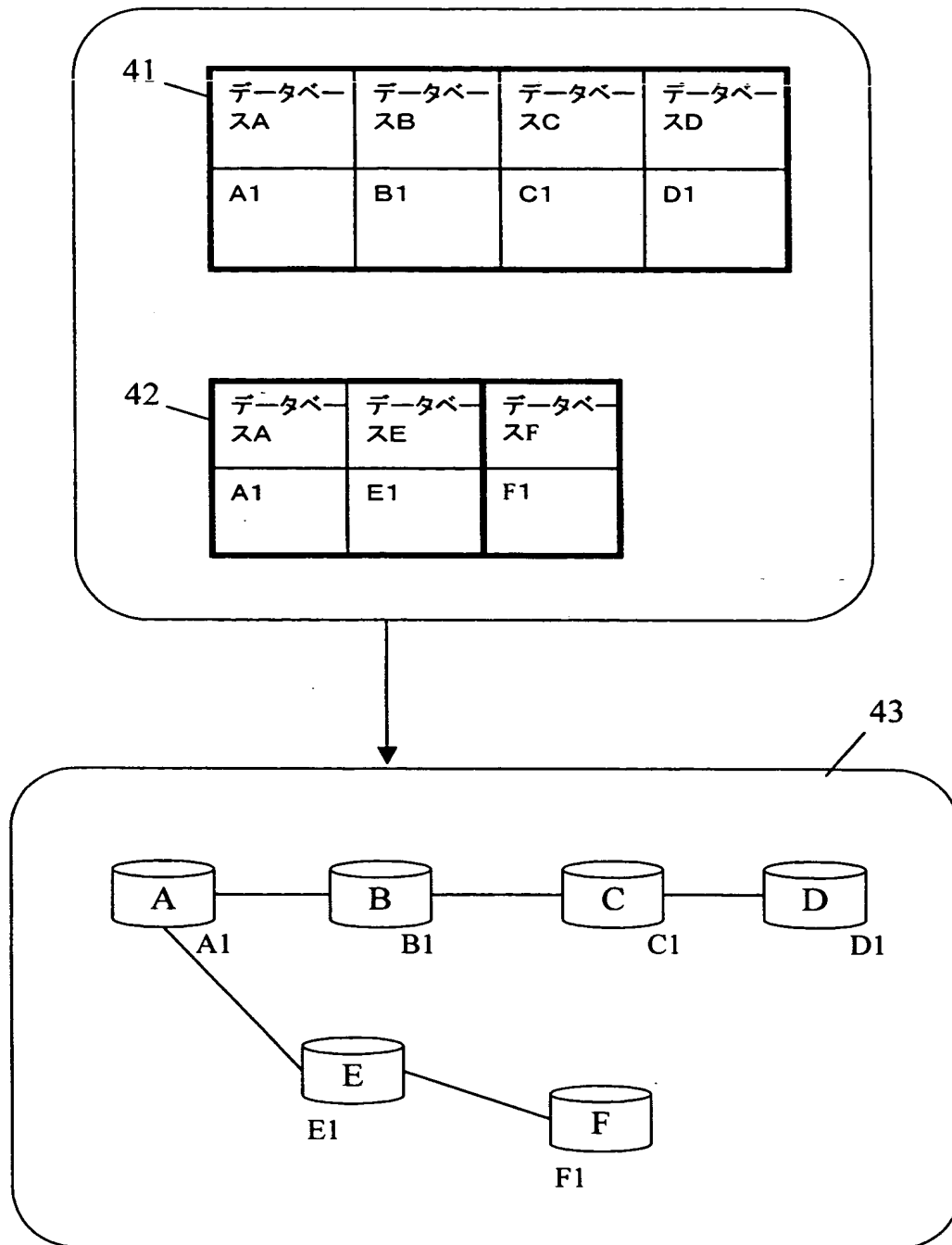
【図 2】



【図 3】



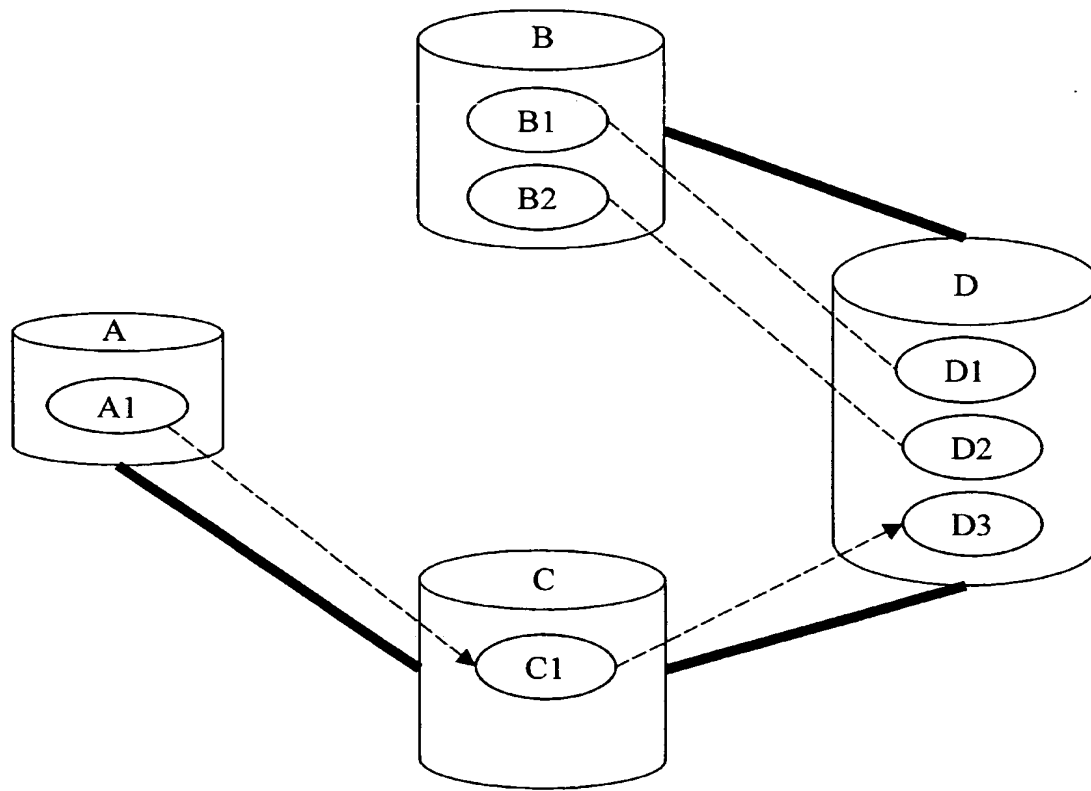
【図 4】



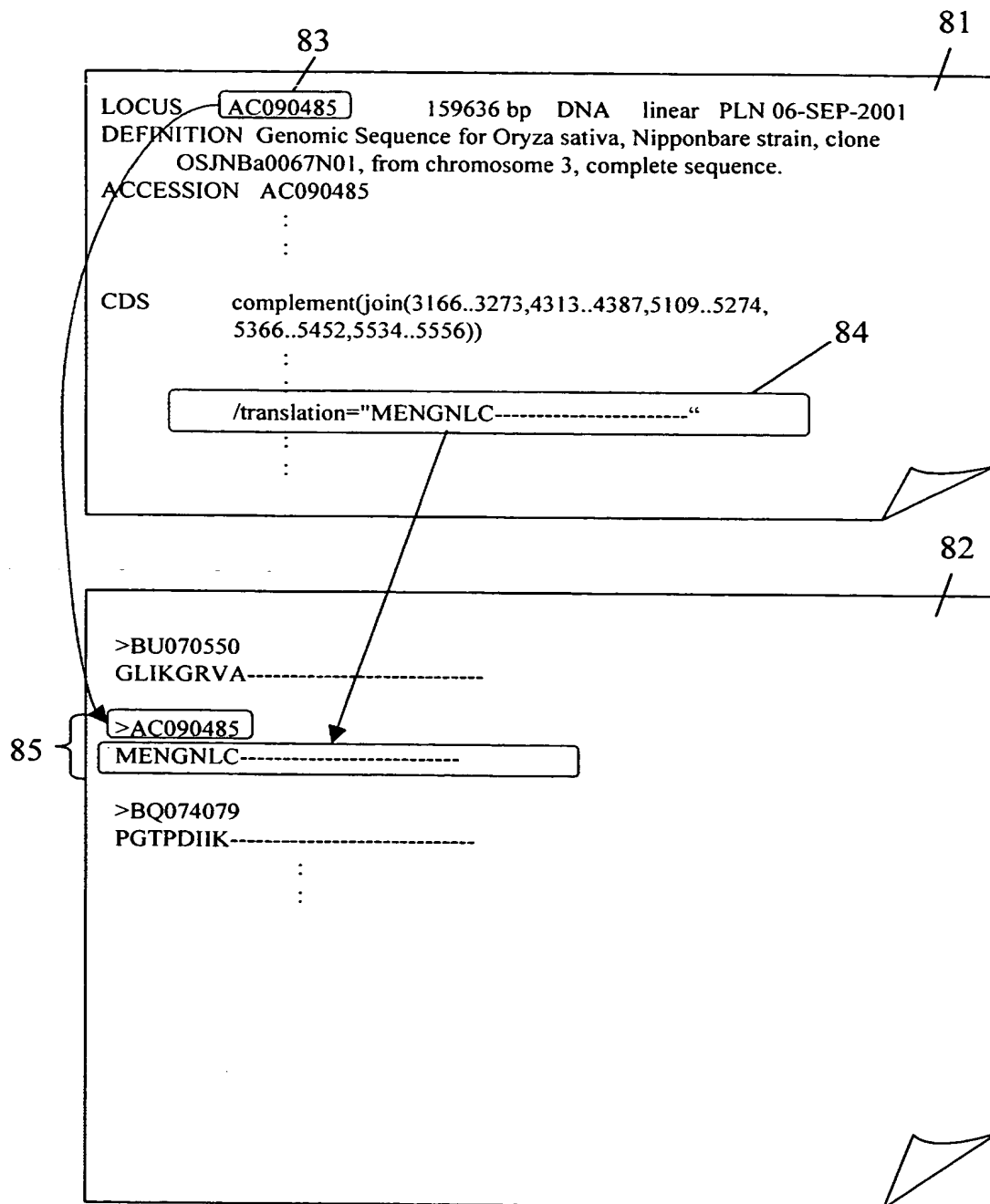
【図 5】

KeyDB	ルート	TargetDB
A	B	C
A	B, C	D
A	B, C	E
....

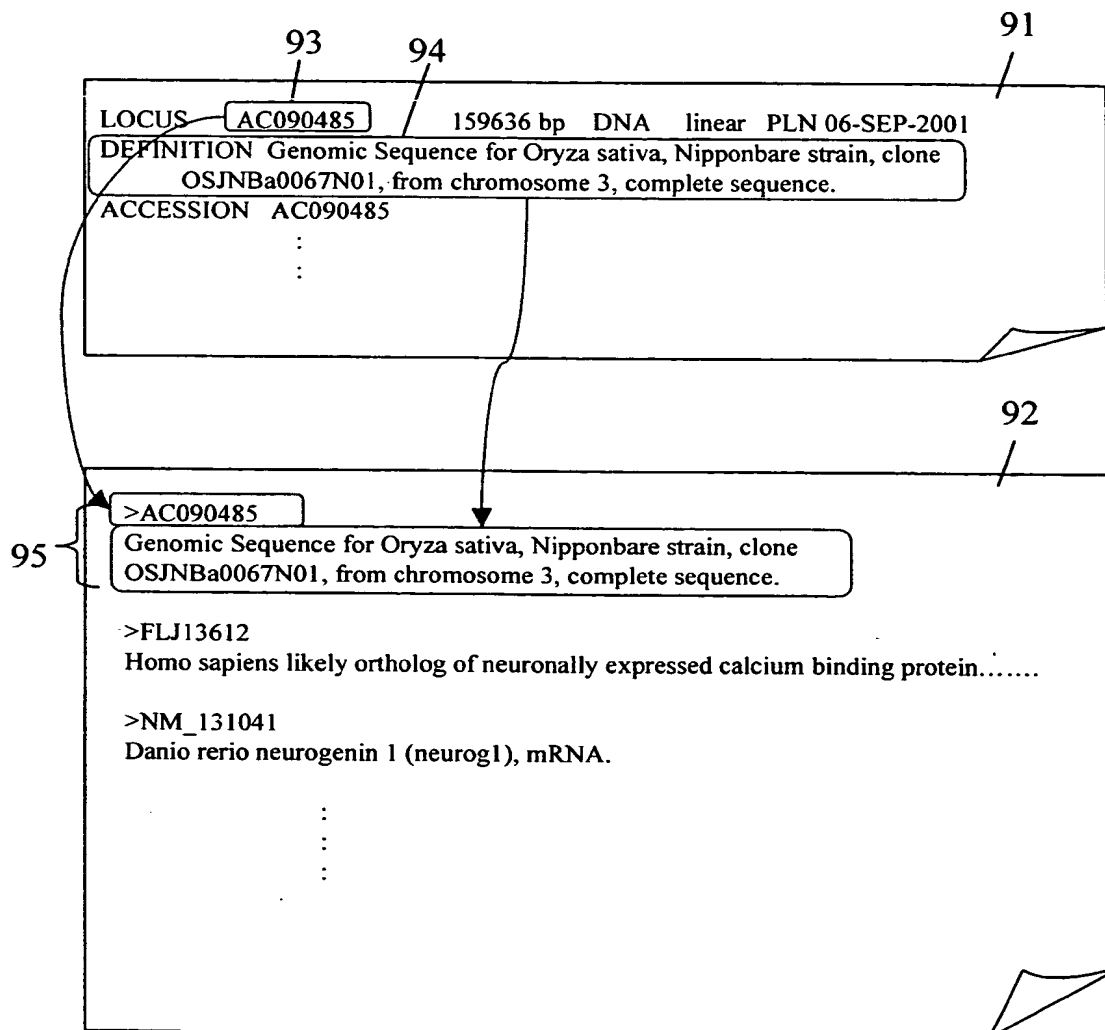
【図 7】



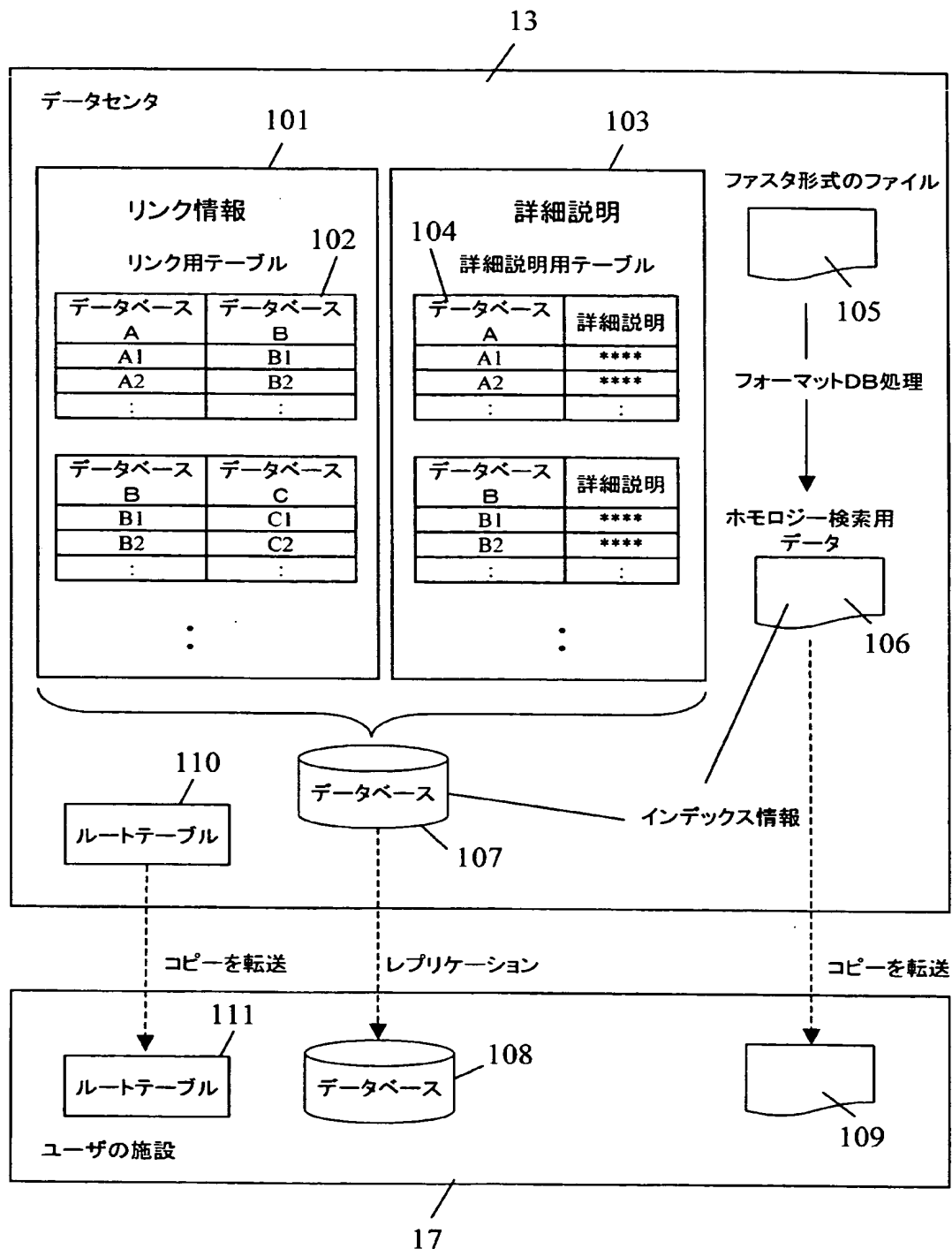
【図 8】



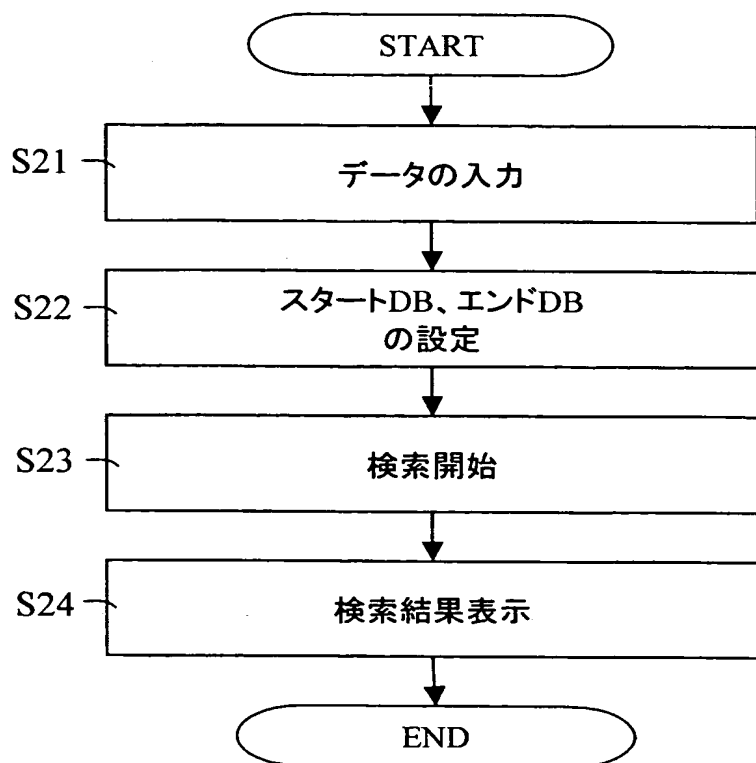
【図 9】



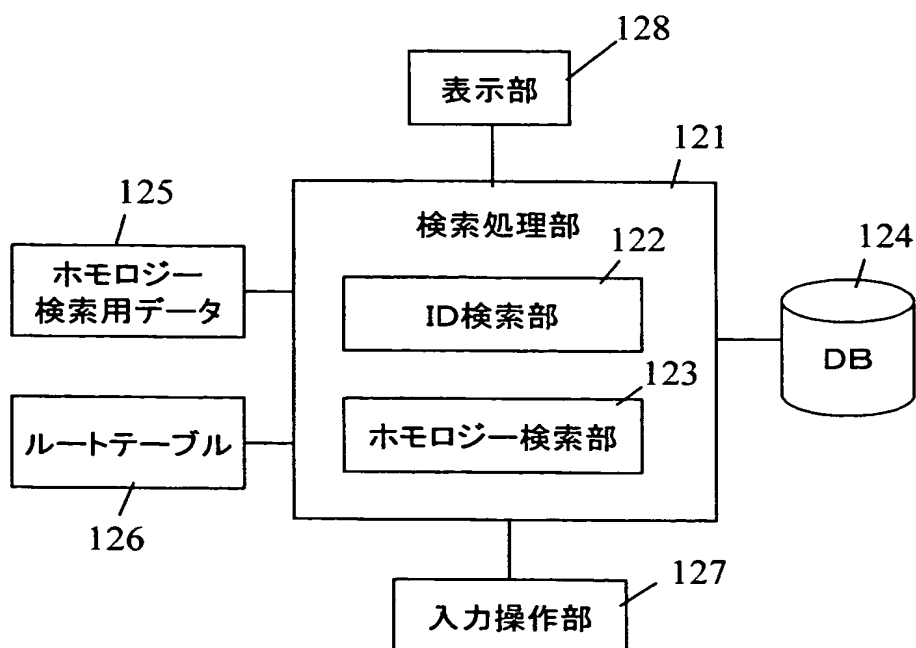
【図 10】



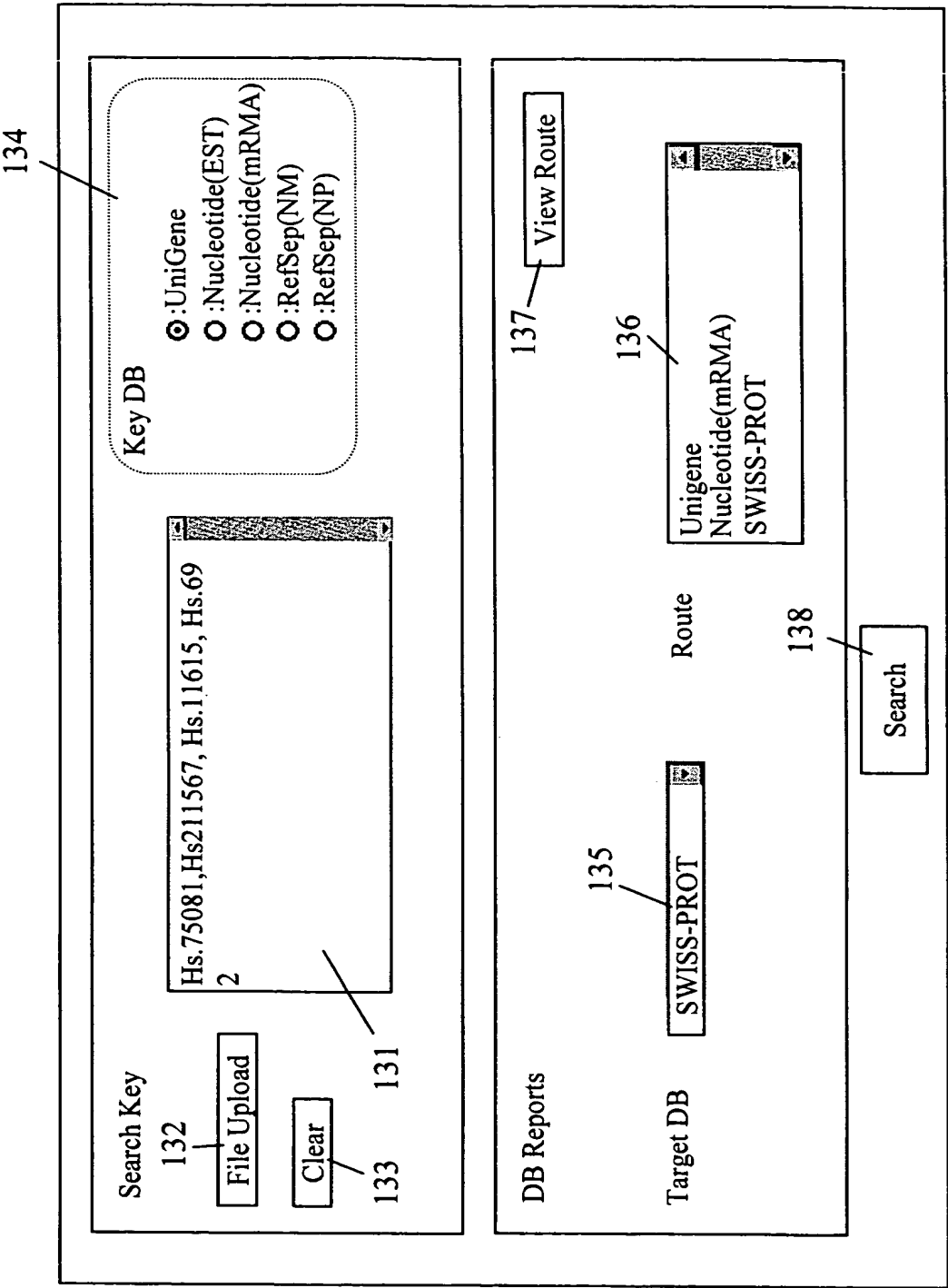
【図 1 1】



【図 1 2】



【図 13】



【図 14】

Query Sequence

File Upload

Clear

>Query001

1 gaattca.....

61

121

181

241

301

361

Program

Blastn(DNA Query vs. DNA DB)

Key DB

Nucleotide(EST)

Detailed Option

Score: Alignments: Except: Word Size: Filter:

10 10 10 10

Other Options:

☒ ON ☐ OFF

DB Reports

Target DB

SWISS-PROT

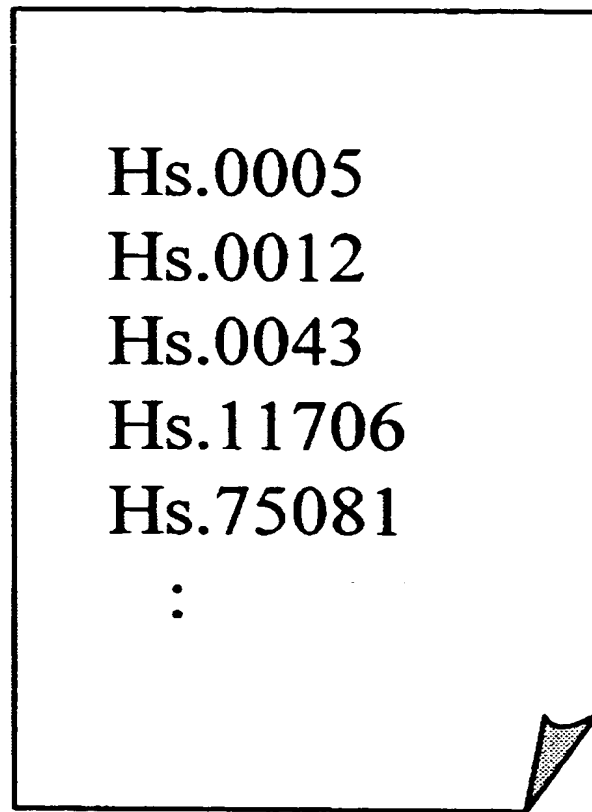
Route

Nucleotide(EST)

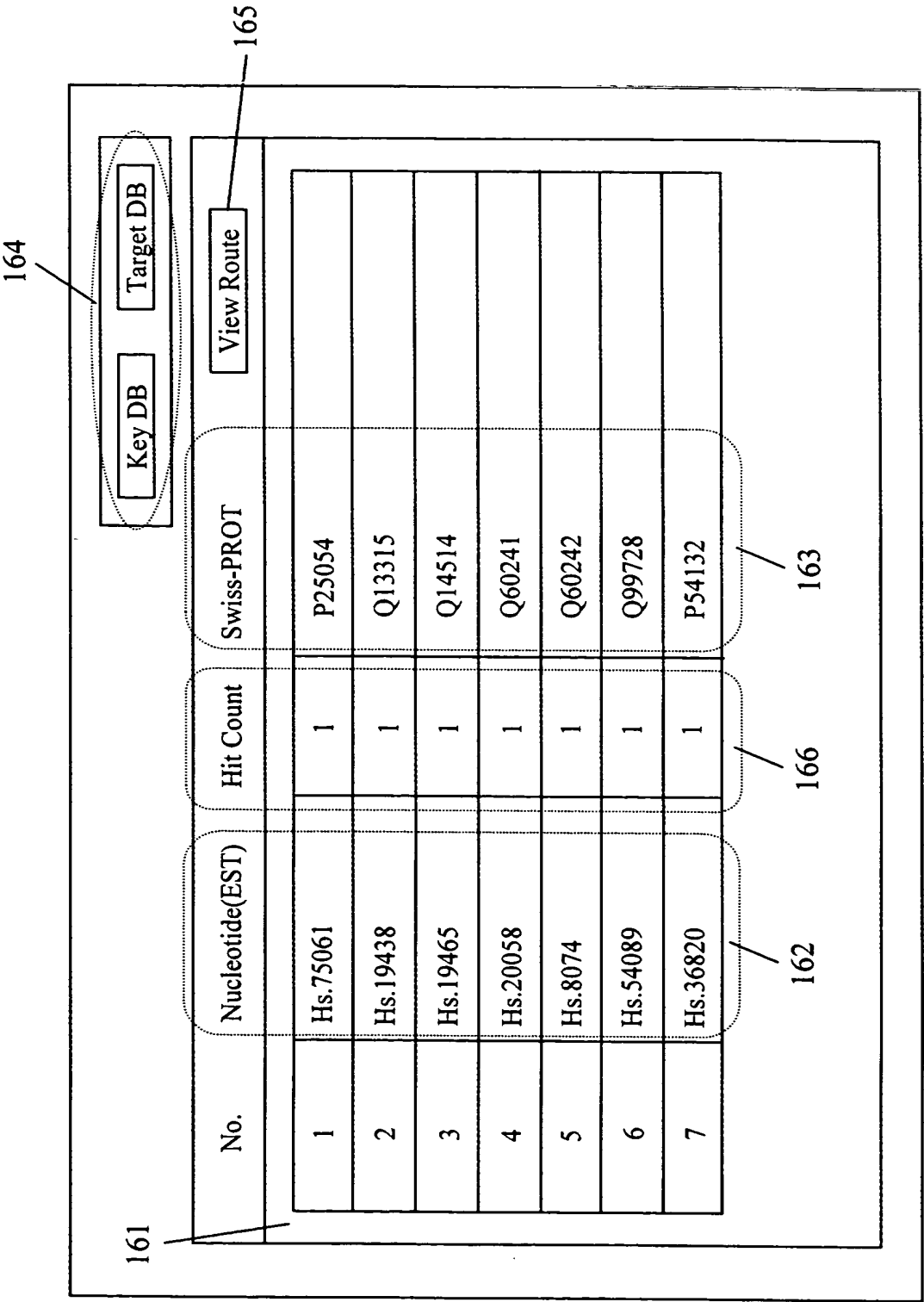
View Route

Search

【図 1 5】



【図 16】



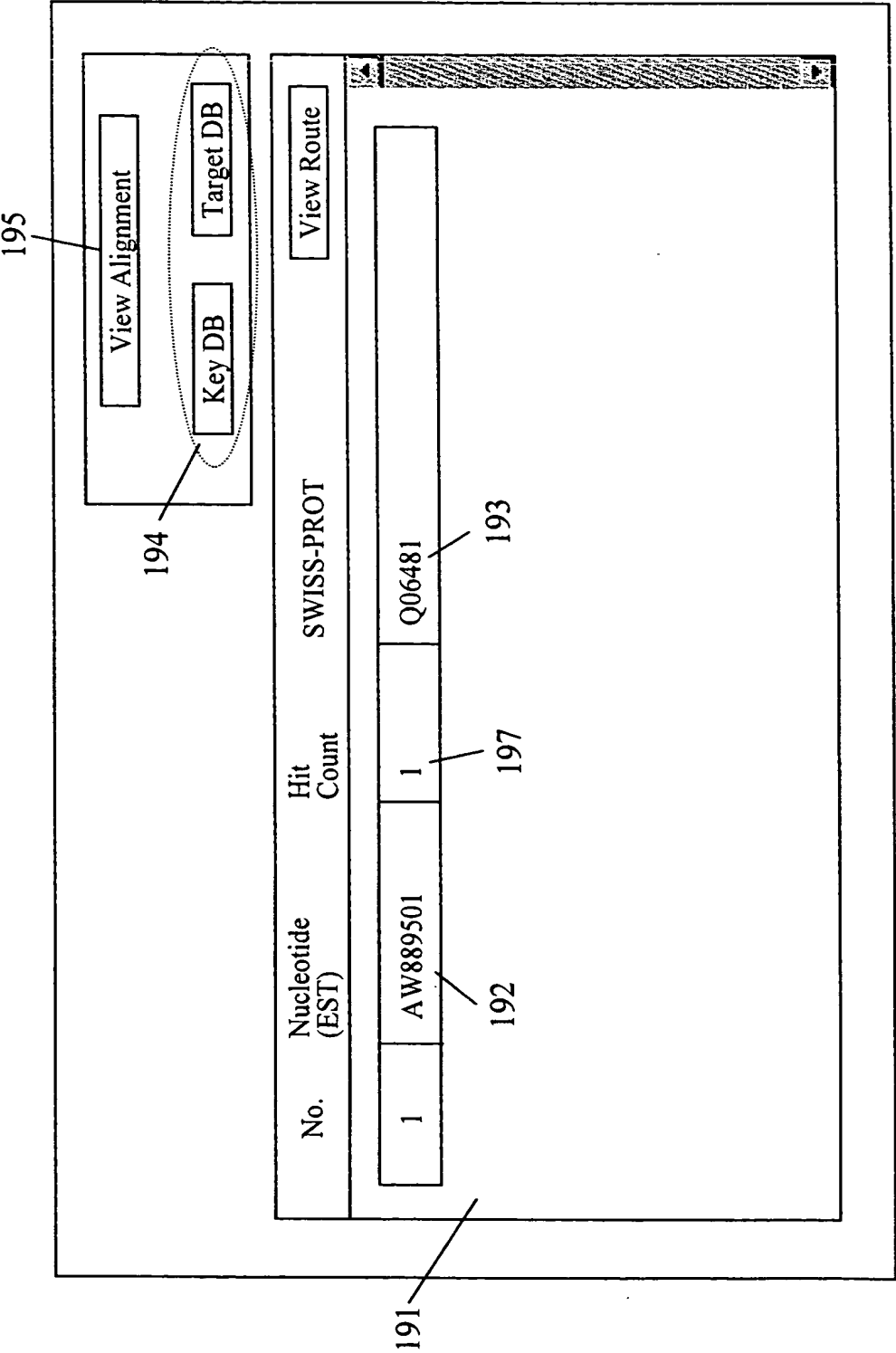
【図 17】

<div>save</div> <div>close</div>		
Database: Nucleotide(EST)		
Search Key	Accession	Definition
Hs.350068	BE385140	601274939F1 NIH_MGC_20 Homo sapiens cDNA clone IMAGE:3615944 5', mRNA sep
Hs.350068	BG758408	602712606F1 NIH_MGC_48 Homo sapiens cDNA clone IMAGE:4853112 5', mRNA sep
Hs.350068	BE275447	601121460F1 NIH_MGC_20 Homo sapiens cDNA clone IMAGE:2988734 5', mRNA sep
Hs.350068	BI834414	603084468F1 NIH_MGC_120 Homo sapiens cDNA clone IMAGE:5223593 5', mRNA sep
Hs.350068	BF205012	60186366F1 NIH_MGC_17 Homo sapiens cDNA clone IMAGE:4106937 5', mRNA sep
Hs.350068	BG750393	602709227F1 NIH_MGC_43 Homo sapiens cDNA clone IMAGE:4845703 5', mRNA sep
Hs.350068	BE260149	601147760F1 NIH_MGC_19 Homo sapiens cDNA clone IMAGE:3613047 5', mRNA sep
Hs.350068	BI758989	603042463F1 NIH_MGC_116 Homo sapiens cDNA clone IMAGE:5183031 5', mRNA sep
Hs.350068	BF027028	601671261F1 NIH_MGC_20 Homo sapiens cDNA clone IMAGE:3954159 5', mRNA sep
Hs.350068	BG749392	602707881F1 NIH_MGC_43 Homo sapiens cDNA clone IMAGE:4844444 5', mRNA sep
Hs.350068	BI910571	603068266F1 NIH_MGC_118 Homo sapiens cDNA clone IMAGE:5216992 5', mRNA sep

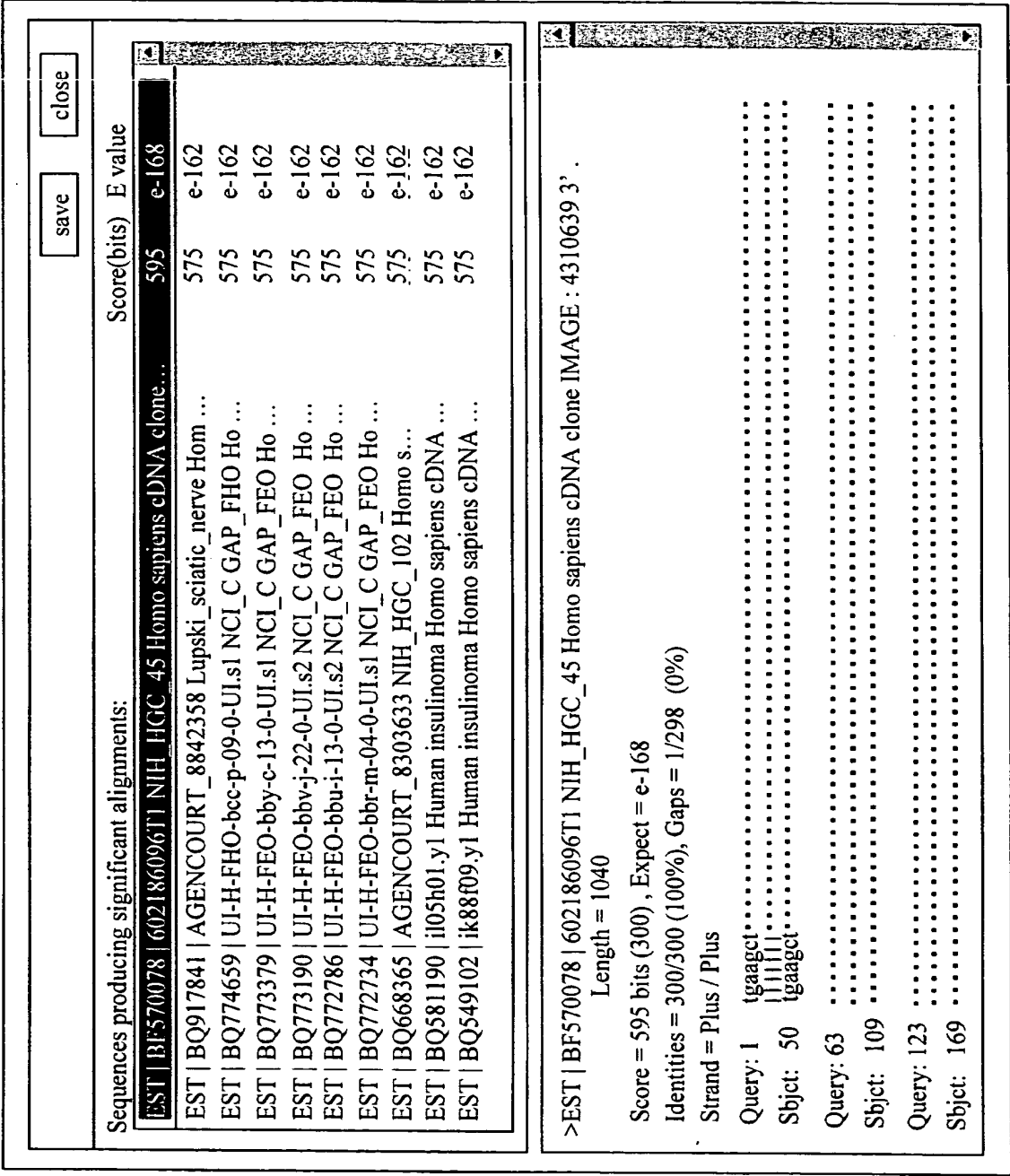
【図 1 8】

> Query001	
1	gaattcca....
61	
121	
181	

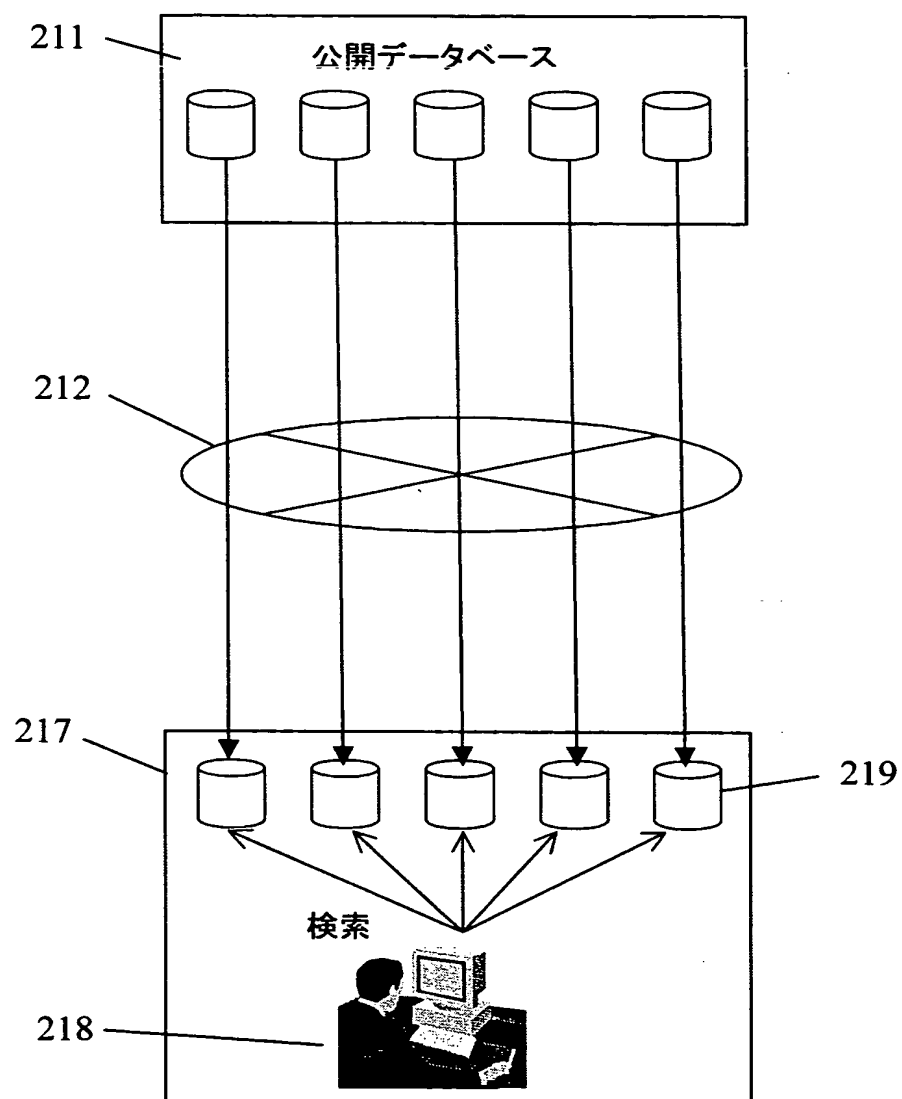
【図 19】



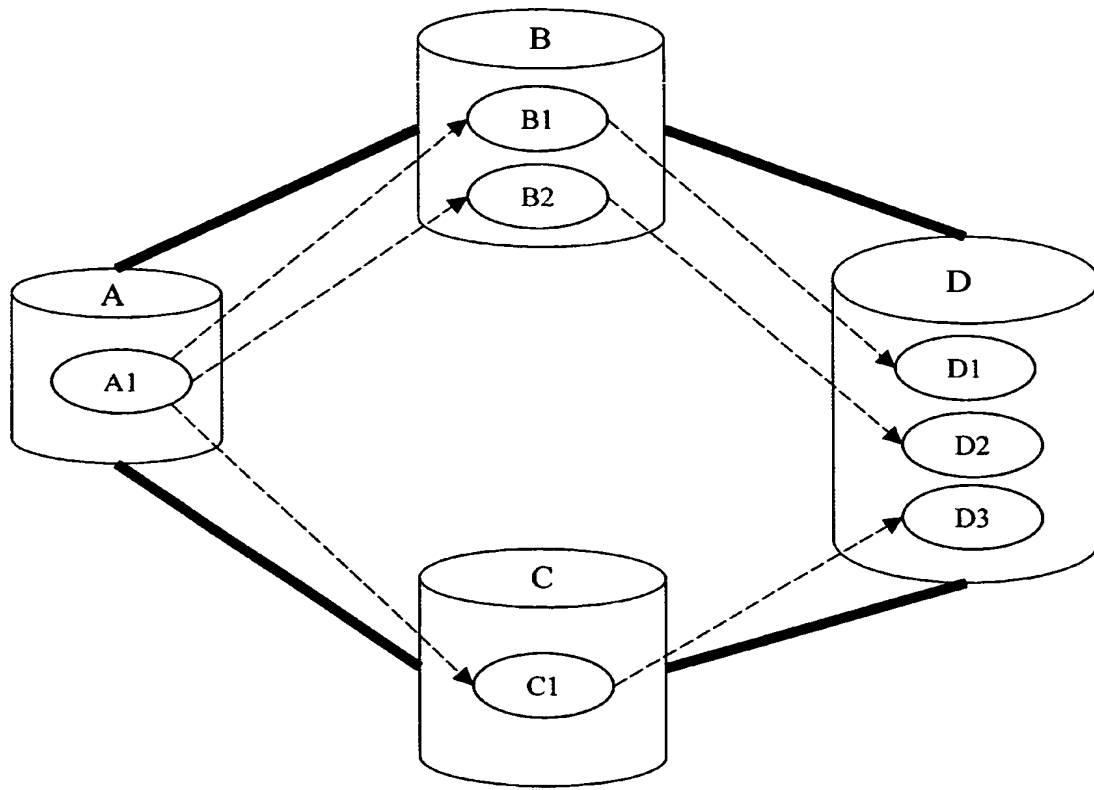
【図 2 0】



【図 21】



【図 22】



【書類名】 要約書

【要約】

【課題】 生体物質情報を格納している複数のデータベースより必要な情報を簡易に抽出する。

【解決手段】 生体物質に関する情報を格納している複数のデータベース 1 1 からデータをデータセンタにダウンロードし、ダウンロードしたデータから、インデックスとして、2つのデータベースのデータ間のリンクを表す情報、各データの詳細説明、及びホモロジー検索用の配列データを抽出し、抽出したインデックス 1 5 をユーザ施設に配信する。ユーザ 1 8 は配信されたインデックスを用いて検索を行う。

【選択図】 図 1

特願 2 0 0 2 - 3 4 4 4 5 2

出 願 人 履 歴 情 報

識別番号

[0 0 0 2 3 3 0 5 5]

1. 変更年月日

1 9 9 0 年 8 月 7 日

[変更理由]

新規登録

住 所

神奈川県横浜市中区尾上町 6 丁目 8 1 番地

氏 名

日立ソフトウェアエンジニアリング株式会社

2. 変更年月日

2 0 0 2 年 1 0 月 1 1 日

[変更理由]

住所変更

住 所

神奈川県横浜市鶴見区末広町一丁目 1 番 4 3

氏 名

日立ソフトウェアエンジニアリング株式会社